DOCUMENT RESUME

ED 204 384 TM 810 374

AUTHOR Ward, Marilyn D.: And Others
TITLE The Observer Effect in Classroom Visitation.
PUB DATE Apr 81
NOTE 21p.: Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classroom Observation Techniques; Faculty Evaluation; Higher Education; *Peer Evaluation; Rating Scales; *Student Evaluation of Teacher Performance; *Teacher Behavior; Test Reliability
IDENTIFIERS Verbal Interaction Category System; Western Illinois Univ Evaluation Questionnaire

ABSTRACT

Using behavioral rating forms, college professors were covertly rated each day by pairs of students ostensibly enrolled in their classes. On specified days pairs of faculty observers introduced themselves to the instructor and rated him using the same forms. The purpose was to determine whether changes occurred in the instructor's behavior as a result of the intrusion of peers into the classroom. Economics, Mathematics, and Psychology classes were observed. Interrater reliabilities were very high. On peer observer days teacher-initiated talk significantly decreased and instructors received significantly higher student ratings than on baseline days. The results of the study have implications for several areas concerning faculty evaluation. (Author/GK)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED204384

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

# The Observer Effect in Classroom Visitation

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M. D. Ward

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Marilyn D. Ward, Western Illinois University

D. Cecil Clark, Brigham Young University

Grant Von Harrison, Brigham Young University

Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, California, April 13-17, 1981

TM 810 374

The crunch of declining enrollments has forced many colleges and universities to take a harder look at the quality of teaching on their campuses. One of the methods which has been very popular as an evaluation of teaching is that of student ratings of faculty.

However, even though the literature is replete with evidence as to the reliability and validity of student ratings (See Guthrie, 1954; Maslow and Zimmerman, 1956; McKeachie, 1969; Costin, Greenough, and Menges, 1971; Swanson and Sisson, 1971; and Menges, 1973), many faculty and administrators are still opposed to the use of student ratings for summative evaluation. One method of evaluating teaching that has been suggested either as an alternative or as an addition to student ratings is colleague evaluation. According to Wilson, Dienst, and Watson (1973) professors commonly assert that only a teacher's colleagues can judge his competency and Aleamoni (1974) observed that many faculty feel that only a colleague is qualified to judge their teaching skills.

There are many indirect measures of instruction which faculty can use to evaluate their colleagues' teaching. Some of those suggested (Wilson, Dienst, and Watson, 1973; Batista, 1976; Centra, 1979; and Cohen and McKeachie, 1980) include: course syllabi, objectives for the course, appropriateness of the textbook, reading lists, nature of assignments, method of evaluating students, and knowledge of the subject matter. One direct measure of instruction frequently mentioned is that of classroom visitation (Bayley, 1967; Miller, 1972; and Berquist, Phillips, and Quehl, 1975). One objection that has been raised about this method of teacher evaluation concerns its reliability. Centra (1975) has concluded that limited classroom observations by peers are not reliable enough to use

for summative evaluation upon which to base decisions about promotion and tenure. Cohen and McKeachie (1980) also stated that classroom observation is probably not a feasible approach for summative teacher evaluation. In spite of these concerns, this method of teacher evaluation continues to be used at many colleges and universities (Gustad, 1967; Seldin, 1975). Astin and Lee (1967) reported that classroom visitation was in use in at least 14 percent of all colleges as a source of information about teaching effectiveness.

The most important drawback to observation or visitation is that the act of visitation itself will almost certainly influence the phenomenon observed. Indeed, according to Gage (1961) an instructor's performance under those circumstances may depend more on nerve than on skill. There is a vast literature (see Barker, 1963; Kintz, Delprato, Mette, Persons, and Schappe, 1965; Sanford, 1965; Webb, Campbell, Schwartz and Sechrest, 1966; Barber, 1968; Jung, 1971; and Rosenthal, 1976) on experimenter effects which offers strong evidence that when an experimenter is present while the subjects are learning a new or different task those subjects are influenced by the mere presence of the experimenter. There is a similar effect noted when an experimenter is present and subjects are doing an already learned task. This is known as the Hawthorne effect. According to Cook (1962) the Hawthorne effect is a phenomenon characterized by an awareness on the part of the subjects as a special treatment created by an artificial experimental condition. Although the possibility of the Hawthorne effect in classroom observation has been acknowledged in the literature (Gage, 1961; Medley and Metzel, 1963; Centra, 1975; and Schwartz, 1980), there have been no empirical studies which have demonstrated a Hawthorne effect in teacher evaluation. In fact, Seldin (1975)

noted that research studies investigating the reliability and validity of classroom visitation for purposes of evaluating colleagues have yet to be carried out. Therefore, it is important to discover what changes occur, if any, in a classroom when an obvious observer is present.

The present study is an attempt to measure changes in instructor behavior caused by the intrusion of visitation teams into the classroom. In order to measure changes in classroom behavior it is necessary to have ongoing measure of classroom behavior against which to measure the change. Of course, the instructor must be unaware of the ongoing measure of classroom behavior so that his behavior will be unaffected by it. In order to obtain such covert measures, students who appeared to be enrolled in the class and were on the instructor's roll were in actuality taking a measure of classroom behavior but appeared to be taking notes. Several times during the first few weeks of the class, faculty visitation teams were also present in the classroom and were visibly taking a measure of classroom interaction. In the absence of any empirical studies, it was impossible to predict for any one instructor if, or what kind of changes in classroom behavior would occur when a peer observation team was present. However, the best possible prediction was that there would be some measurable differences in classroom behavior when obvious observers were present as opposed to when they were not. Among possible behaviors we thought might change was that an instructor would try to involve the students more if only by asking more questions. We also speculated that an instructor might try to use more humor, act more interested in the class in general and the students individually, and generally make an attempt to be a "better teacher."

Any sort of deception raises the issue of whether or not it is ethical to conduct an experiment without a subject's full knowledge. Berg (1954) flatly stated that it is not moral to do so. However, most views are that informed consent, to the extent practical, along with full disclosure in a thorough debriefing session will serve in those situations where the experiment presents no physical or psychological danger to the subject. In evaluating teaching through classroom visits, the ethical question might not seem so serious a consideration if the instructors would agree to be observed without specification as to when or by whom the observation would take place.

The following study is an attempt to measure the magnitude of the observer effect caused by classroom visitation while meeting the ethical concerns that arise from this sort of experiment.

## Method

Subjects. A total of six male instructors from Western Illinois University served as subjects in this experiment. All were at the rank of Assistant Professor or above, all had at least eight years experience teaching in higher education, and all were tenured. Instructors were randomly chosen from the spring semester class schedule with the restrictions that one instructor from an introductory course and one instructor from an upper division course be included from each of the following departments: Economics, Mathematics, and Psychology. Classes taught by these instructors are shown in Table 1.

Covert Student Raters. A total of 12 students were hired to register for and attend the first 10 days (approximately 3½ weeks) of regular

class meetings and to rate instructors during and immediately after each class session. These students were volunteers who were chosen to participate based upon their availability for training during semester break. Raters were randomly assigned to the six instructors, two per instructor.

Overt Faculty Raters. Thirty-six faculty members were asked to participate in this experiment. They were asked on the basis that all had shown some interest in evaluation of teaching as evidenced by their participation in faculty development workshops on that subject. Faculty raters were from diverse disciplines including education, psychology, physics, business, physical education, health, home economics, history, music, speech, mathematics, english, and geography. The two raters in a pair were always from different disciplines and always observed a class not of their own discipline. Given the restrictions stated above, the overt faculty raters were randomly assigned to the instructors, two per instructor per class session. Each pair of overt faculty raters attended only one class period.

Evaluation Instruments. In order to try and measure any finite changes that might take place in an instructor's behavior, several measures were taken, two of which are reported here. The Western Illinois University Evaluation Questionnaire (WIUEQ) was used to evaluate student impressions of the instructor. The WIUEQ was adapted from the University of Illinois Course Evaluation Questionnaire (CEQ) which has been used extensively as a student rating form (Aleamoni and Spencer, 1973; Aleamoni, 1978). The CEQ has perhaps the most extensive reliability, validity, and norm base data of all student rating forms (Aleamoni, 1978) as normative data has been collected since 1966. The CEQ is used to register student

attitudes toward a course and an instructor. Items on the WIUEQ that refer specifically to the course such as questions about the difficulty of the course material or the course content seemed inappropriate to use as a measure of instructor performance. Therefore, a specific item "The instructor was excellent" and a combination of items used to rate the instructor generally were chosen to measure the students' impressions of the instructor on a day-to-day basis. The computer scoring for the selected item and the combined instructor rating are in terms of percentiles with all instructors at WIU as a comparison base.

The Verbal Interaction Category System (VICS) (Amidon and Hunter, 1966) was used as an instrument to evaluate teacher-student interactions. The VICS is based upon the Flanders system of Interaction Analysis, and contains 17 types of behaviors grouped together into five major categories for analyzing classroom verbal behavior. These are: teacher initiated talk, teacher response, pupil response, pupil initiated talk, and other.

Procedure. Instructors were asked to sign a form indicating their willingness to have their classes visited and evaluated during the first few weeks of the semester as part of a study of the use of a rating form. They were assured that the results of their ratings would be anonymous, i.e., their name would not be attached to the ratings and the ratings would be used only for comparison with other teachers and not used to evaluate the instructors for merit, promotion, tenure, etc. All instructors who were asked agreed to participate in the study.

Student raters were trained in the use of VICS rating scale using videotapes until satisfactory interrater reliability was achieved ($r \geq .92$). Overt faculty raters were sent written instructions concerning their visit to the classroom. These directions instructed them to arrive early

enough to introduce themselves to the instructor. They were also sent written directions for using the VICS to assess classroom interaction.

Student raters were given their assigned classes and were registered for the courses as regular students. They evaluated their instructors beginning on the second day of class and every class day for 10 regular class meetings of the semester and then withdrew from the class. On each regular class day, student raters evaluated the class using the VICS evaluation instrument under the guise of notetaking. Immediately after each regular class period student raters filled out the entire WIUEQ. The ratings were coded only with a number and student raters mailed all data from each observation through campus mail to the experimenter as soon as possible upon leaving the class observed.

On the 4th, 7th, and 10th regular class day, pairs of overt faculty raters attended each of the six classes in the study. Each overt faculty rater pair attended only one class period. Covert student raters were present and a given pair of covert student raters remained in the same class for the duration of the experiment. Baseline data was collected on the 1st, 2nd, 3rd, 5th, 6th, 8th, and 9th regular class days of the study and experimental data was collected on the 4th, 7th, and 10th regular class days of the study which were the three class sessions for each instructor which were attended by the overt faculty rater pairs. If the instructor was giving a test, or showing a movie, the covert student raters and the overt faculty raters were instructed to leave and that day was not included in the study. Each pair of overt faculty raters only observed a single class session.

Debriefing. At the conclusion of the experiment, the experimenter met individually with each of the six instructors, explained the nature

of the study, described the results, and assured them that their names would not be attached to the ratings. The experimenter also invited the peer observers to a debriefing session and explained the nature of the study and presented the results to them.

## Results

WIUEQ. The mean score on observer days was compared to the mean scores on the baseline days for the item "The instructor was excellent." The t-test was significant ($t=3.26$, $df=5$, $p<.05$) with the mean on observer days being at the 42.23 percentile and the mean on baseline days being at the 34.37 percentile. Means on baseline days and observer days for each instructor and for all instructors combined are shown in Figure 1. It can be seen that instructors on the average were able to improve their rating by almost a decile from baseline days to observer days.

Another measure of instructor rating on the WIUEQ combined several items into one general instructor rating. The mean score on observer days was compared to the mean scores on the baseline days by means of a paired observations t-test. The t-test was significant ($t=7.38$, $df=5$, $p<.01$) with the mean on the observer days being at the 50th percentile and the mean on the baseline days being at the 43.22 percentile. Means on baseline days and observer days for each instructor and for all instructors combined are shown in Figure 2. It can be seen that professors on the average were able to improve their ratings from the 43rd percentile on baseline days to the 50th percentile on observer days.

Inspection of the data in both Figure 1 and Figure 2 fail to reveal any differential effects between upper and lower division classes, or for class size.

VICS. It quickly became apparent in training student raters to use the VICS that the majority of classroom interaction consisted of teacher-initiated talk. Therefore, the median percentage of teacher-initiated talk for each quarter period of baseline days and observer days for each instructor served as the data for this measure. These measures for each instructor and for all instructors combined are shown in Figure 3. These data were subjected to a 2 x 4 repeated measures factorial analysis of variance incorporating time (measures taken each quarter-hour) and type of observation (baseline and observer). The results were significant for both time ($F=5.06$, $df=3/15$, $p<.05$) and type of observation ($F=15.93$, $df=1/5$, $p<.02$). However, the interaction was not significant ($F=.90$, $df=3/15$, $p>.25$). The significant main effect for type of observation showed a lower percentage of teacher-initiated talk on observer days and higher percentage of teacher-initiated talk on baseline days. The mean median percentage of teacher-initiated talk on observer days was 74.24 and the mean median percentage of teacher-initiated talk on baseline days was 79.96. The significant main effect for time showed less teacher-initiated talk at the beginning and end of the period and more in the middle of the total time period. This is not surprising since teachers might be more likely to entertain student questions both at the start and at the end of class periods. Both effects seem strong for some individual instructors and much less for others. For one instructor (EU) the type of observer effect was not very strong while for two other instructors (BU and MU) that effect was particularly noticeable. It was not surprising that this effect varies among instructors but it is interesting to note that the effect is present to some degree for all instructors.

## Conclusions

The results of this study have demonstrated a strong observer effect in classroom visitation. Student ratings of faculty significantly improved on those days when peer observers were present. Measures of the instructors behavior showed that teacher-initiated talk significantly decreased when peer observers were present.

This observer effect should be the weakest possible observer effect because: (1) the instructors were told that this was a research study and that any data obtained would not be attached to them by name, (2) the observations were not scheduled but were unannounced, and (3) the peer observers were all from outside the instructor's discipline.

It can be assumed that knowing that the data from an observation will be used for purposes of evaluation for promotion and/or tenure would increase this effect. In this instance while instructors might try to appear as "good teachers" the demand characteristic would not be as severe as when being evaluated in a tenure or promotion situation.

The results of this study have implications for several areas concerning evaluation of faculty. Since student ratings of instructors changed almost a decile when unannounced observers were present it is possible that instructors could change even more if they knew ahead of time what day the observers would be there. Several articles (Hoyt and Howard, 1978; Gray, 1969; Medley and Mitzel, 1963) have purposed peer evaluation systems that incorporate scheduled classroom visits. It appears that with chances to prepare not only the lecture but perhaps visual aids and jokes that a teacher might further change his student ratings to an even more favorable number. Ratings obtained by peer

observation, especially those with scheduled visits, should be viewed with caution in light of the results of this experiment.

It is also possible that if the observers had been from within the instructor's discipline or particularly if the observers had been Deans or Department Chairmen that the instructor's efforts to improve his teaching would be intensified. It seems reasonable to assume that observers who would not criticize what was being taught would not be as threatening as colleagues from within one's own discipline.

The results of the present study also bear upon the case of student evaluations. Several studies (Miller, 1971; Centra, 1973) have reported that student evaluations do not vary significantly over the semester with the implication that student perceptions of instructors do not change. The study demonstrates that student impressions can vary dependent upon the instructor's behavior on a given day. Of course it should be recalled that in this study the rating given by the students concerned only items dealing with the instructor. Student rating forms that use items exclusively directed toward the course content, examinations, textbooks, use of media, etc. would not be expected to vary in the same way.

Many faculty development programs use formative evaluation with peer evaluation as a component. It may not be prudent to assume that ratings made by peer observers are a true reflection of the instructor's usual teaching. If not, then peer observation may lead to less helpful formative evaluation.

Table 1

Classes Observed

| Department | Level | Code | Size (Approx.) |
|---|---|---|---|
| Psychology | Lower Division | PL | 200 |
| Psychology | Upper Division | PU | 30 |
| Mathematics | Lower Division | ML | 180 |
| Mathematics | Upper Division | MU | 30 |
| Economics | Lower Division | EL | 55 |
| Economics | Upper Division | EU | 35 |

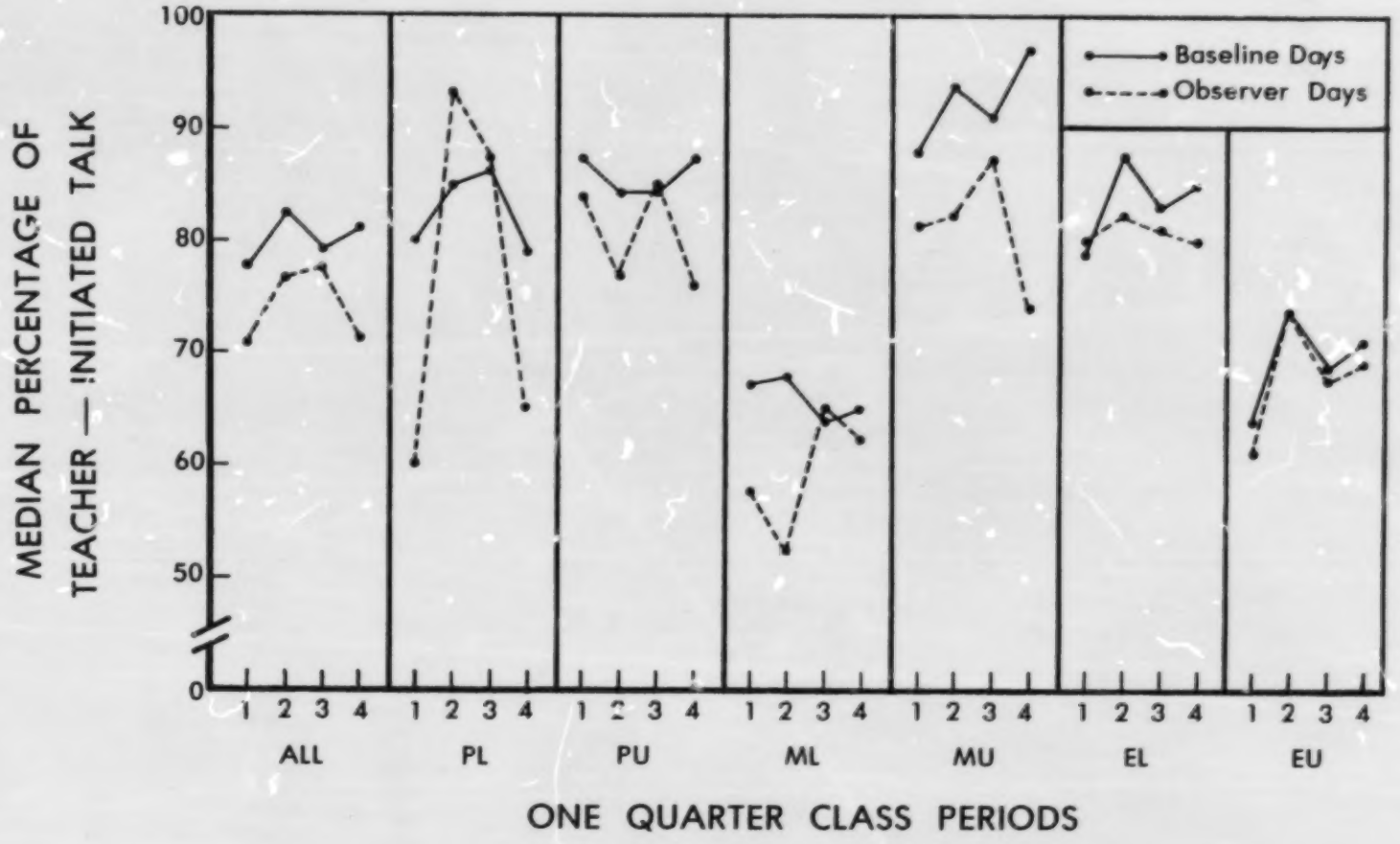




*Figure 3.* Median percentage of teacher - initiated talk for each quarter period for each instructor and mean median percentage of teacher - initiated talk for each quarter period for all instructors (see table 1 for instructor code)

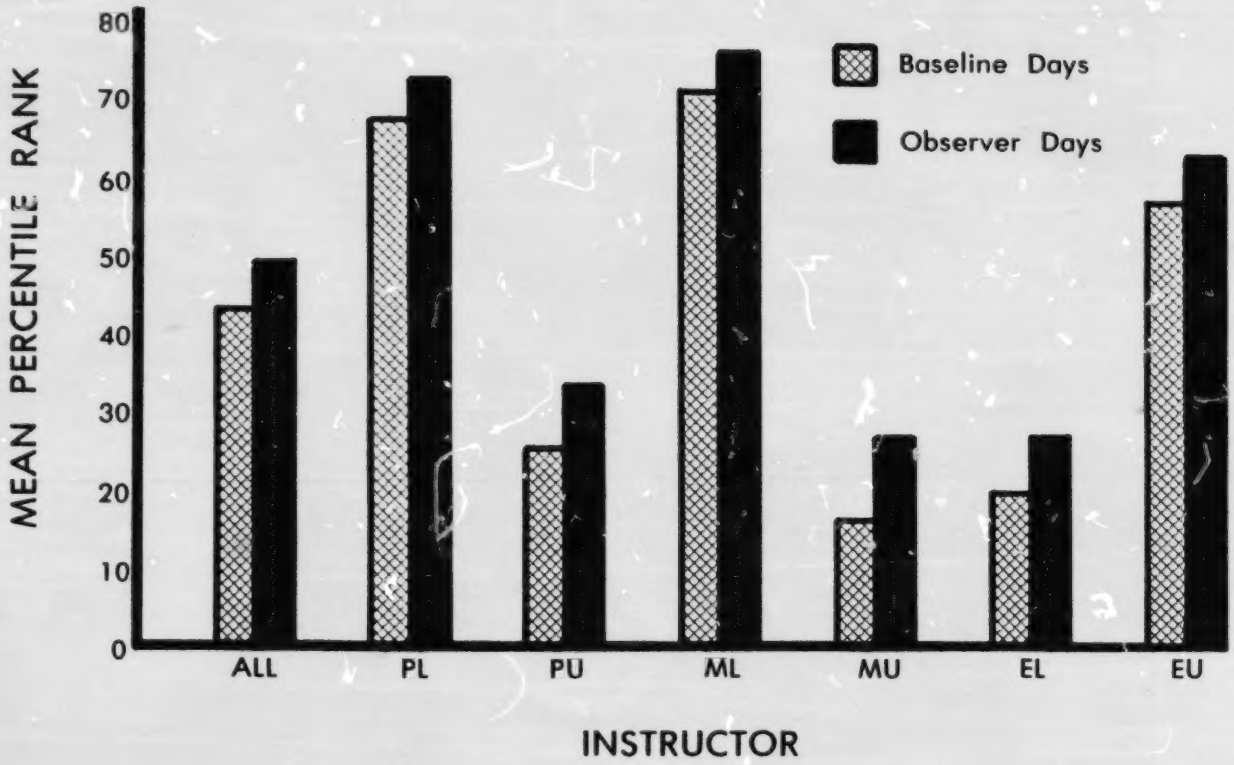


*Figure 2.* Mean rating ( percentile rank ) for combined "Instructor - general" items (see table 1 for instructor code)

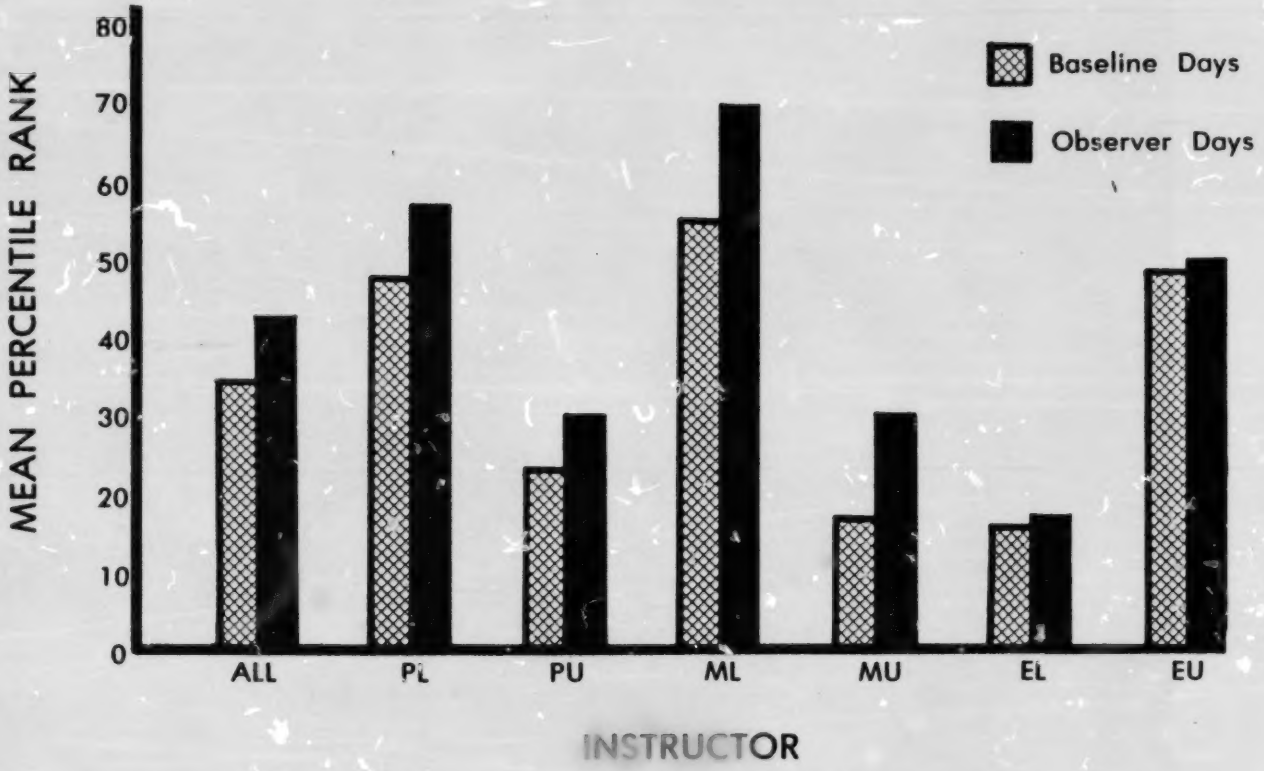


*Figure 1.* Mean rating ( percentile rank ) for "The instructor was excellent" ( see table 1 for instructor code )

References

Aleamoni, L. M. & Spencer, R. E. The Illinois course evaluation questionnaire: A description of its development and a report of some of its results. Educational and Psychological Measurement, 1973, 33, 669-684.

Aleamoni, L. M. Typical faculty concerns about student evaluation of instruction. Paper presented at the Symposium on Methods of Improving University Teaching, 1974, Haifa, Israel.

Aleamoni, L. M. The usefulness of student evaluations in improving college teaching. Instructional Science, 1978, 7, 95-105.

Amidon, E. & Hunter, E. Improving teaching. Holt, Rinehart and Winston, Inc.: New York, 1966.

Astin, A. A. & Lee, C. B. T. Current practices in the evaluation and training of college teachers. In Lee, C. B. T. (Ed.) Improving college teaching. Washington, D. C.: American Council on Education, 1967.

Barber, T. X. & Silver, M. J. Fact, fiction, and the experimenter bias effect. Psychology Bulletin Monographs, 1968, 70, 1-29.

Barker, R. (Ed.) The stream of behavior. New York: Appleton-Century-Crofts, 1963.

Batista, E. E. The place of colleague evaluation in the appraisal of college teaching: A review of the literature. Research in Higher Education, 1976, 4, 257-271.

Bayley, D. H. Making college teaching a profession. Improving College and University Teaching, 1967, 15, 115-119.

Berg, A. The use of human subjects in psychological research. The American psychologist, 1954, 9, 108-111.

Berquist, W. H., Phillips, S. and Quehl, G. H. A handbook for faculty development. Dansville, New York: Dansville Press, Inc., 1975.

Centra, J. A. Colleagues as raters of instruction. Journal of Higher Education, 1975, 46, 327-337.

Centra, J. A. Determining faculty effectiveness. San Francisco: Jossey-Bass, 1979.

Centra, J. A. Effectiveness of student feedback in modifying college instruction. Journal of Educational Psychology, 1973, 65, 395-401.

Cohen, P. A. & McKeachie, W. J. The role of colleagues in the evaluation of college teaching. Improving College and University Teaching, 1980, 28, 147-154.

Cook, D. L. The Hawthorne effect in educational research. Phi Delta Kappan, 1962, 44, 116-122.

Costin, F., Greenough, W. T., & Menges, R. J. Student ratings of college teaching reliability, validity, and usefulness. Review of educational research, 1971, 19, 511-535.

Gage, N. L. The appraisal of college teaching. Journal of Higher Education, 1961, 32, 17-22.

Gray, C. E. The teaching model and evaluation of teaching performance. Journal of Higher Education, 1969, 40, 636-640.

Gustad, J. W. Evaluation of teaching performances: Issues and possibilities. In Lee, C. B. T. (Ed.) Improving College Teaching. Washington, D. C.: American Council on Education, 1967.

Guthrie, E. R. The evaluation of teaching: A progress report. Seattle: University of Washington Press, 1954.

Hoyt, D. P. & Howard, G. S. The evaluation of faculty development programs. Research in Higher Education, 1978, 8, 25-38.

Jung, J. The experimenter's dilemma. New York: Harper and Row, 1971.

Kintz, B. L., Delprato, D. J., Mettee, D. R., Persons, C. E., & Schappe, R. H. The experimenter effect. Psychology Bulletin, 1965, 63, 223-232.

McKeachie, W. J. Student ratings of faculty. American association of university professors bulletin, 1969, 55, 439-444.

Maslow, A. H. & Zimmerman, W. College teaching ability, scholarly activity, and personality. Journal of educational psychology, 1956, 47, 185-189.

Medley, D. M. & Mitzel, H. E. Measuring classroom behavior by systematic observation. In Gage, N. L. (Ed.) Handbook of Research on Teaching. Chicago: Rand McNally, 1963.

Megan, N. The dynamics of evaluation. In Lee, C. B. T. (Ed.) Improving college teaching, Washington, D. C.: American Council on Education, 1967.

Menges, R. J. The reporters: Students rate instruction. In Pace, C. R. (Ed.) Evaluating teaching and learning. San Francisco: Josey-Bass, 1973.

Miller, M. T. Instructor attitudes toward, and their use of student ratings of teachers. Journal of Educational Psychology, 1971, 62, 235-239.

Rosenthal, R. Experimenter Effects in Behavioral Research. New York: Irvington Publishers, Inc., 1976.

Sanford, N. Will psychologists study human problems? American Psychology, 1965, 20, 192-202.

Schwartz, L. L. Criteria for effective university teaching. Improving College and University Teaching, 1980, 28, 120-123.

Seldin, P. How colleges evaluate professors. New York: Blythe-Pennington, 1975.

Swanson, R. A. & Sisson, D. J. The development, evaluation, and utilization of a departmental faculty appraisal system. Journal of industrial teacher education, 1971, 9, 64-79.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. Unobtrusive measures: Nonreactive research in the social sciences. Chicago: Rand McNally, 1966.

Wikon, R. C., Dienst, E. R. & Watson, N. L. Characteristics of effective college teachers as perceived by their colleagues. Journal of Educational Measurement, 1973, 10, 31-37.